**Speaker-Invariant Emotion Recognition with Adversarial Learning**

**Student:** Bryan Leow Xuan Zhen

**Supervisor:** Professor Jagath C Rajapakse

**Motivation**
Recent advances in technology have given birth to intelligent speech assistants such as Alexa and Siri which can perform a myriad of tasks just from the end users’ voice command. However, they still lack the capability to recognize human emotions when formulating a response. For such speech assistants to be useful to the general population, not only is it important for the underlying Speech Emotion Recognition system be able to recognise human emotions, but the representation captured should also be speaker-invariant.

**Methodology**

**Audio Signal Processing**
1. Silence and background noise from the raw audio recording is removed using librosa and scipy package.
2. To perform batch training, every input must be of the same size. Hence, the audio recordings are padded to the longest audio length in the dataset.
3. The pre-processed audio recordings are then converted to Mel Frequency Cepstral Coefficients (MFCCs) before feeding into the encoder part of our model architecture.

![Conversion of a pre-processed audio recording to its MFCCs](image)

**Model Architecture**
Our model architecture is composed of 3 components:

- i. An encoder made from 2D Convolutional Neural Network (CNN) and Bi-directional Gated Recurrent Unit (biGRU) layers and a statistical pooling layer. Together, they generate speaker invariant representations.
- ii. An emotional classifier that predicts emotional labels.
- iii. A speaker classifier which removes speaker variability, hence making the entire training an adversarial learning.

![Model Architecture Diagram](image)

**Experiments and Results**
Our experiments are conducted on the EmoDB and RAVDESS dataset using K-fold leave-two-speakers-out cross-validation and testing. In each fold, 2 speakers are used for validation and 2 speakers used for testing, with the rest of the speakers used for training. The ratio of male to female speakers in training, validation and testing were set to 1:1.

We experimented with different number of CNN layers for our 2D CNN biGRU encoder. For every depth, we compare the performance of our 2D CNN biGRU encoder with different value of γ where a higher γ value correspond to a higher rate and degree of speaker adversarial training.

To understand how our 2D CNN biGRU encoder perform relative to other encoders utilising adversarial learning in speaker independent emotion recognition tasks, we perform the same K-fold leave-two-speakers-out cross-validation and testing on those encoders. Our experiments show that our proposed architecture achieves the best performance compared to existing models on both EmoDB and RAVDESS dataset.

<table>
<thead>
<tr>
<th>Input features</th>
<th>Encoder</th>
<th>EmoDB</th>
<th>RAVDESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogMF, energy, pitch, MFCC</td>
<td>TDNN BiLSTM</td>
<td>32.4±0.9</td>
<td>44.2±1.7</td>
</tr>
<tr>
<td></td>
<td>1D CNN BiGRU</td>
<td>2D CNN biGRU</td>
<td>73.6±7.1</td>
</tr>
<tr>
<td></td>
<td>61.2±8.4</td>
<td>54.2±8.7</td>
<td>51.3±11.1</td>
</tr>
</tbody>
</table>

**Web Application for Demonstration**

The model shown in this presentation is trained with 5-fold leave-two-speakers-out cross-validation on RAVDESS dataset.